ÖREBRO UNIVERSITY
SWEDISH BUSINESS SCHOOL

WORKING PAPER

14/2009 (revised)

Mitigating Hypothetical Bias in Value of Time Studies: Lab-Experiment Results

Lars Hultkrantz and Shengcong Xue

DEPARTMENT OF ECONOMICS

ISSN 1403-0586

Revised for Journal of Choice Modeling, Oct 30, 2010

Mitigating Hypothetical Bias in Value of Time Studies: Lab-Experiment Results

Lars Hultkrantz* and Shengcong Xue

Swedish Business School, Örebro University

Abstract: We present results from a series of willingness-to-accept value-of-time choice experiments with students in Sweden and China, using both real and hypothetical purchases of the students´ time. Our results confirm negative hypothetical bias in stated choice elicitation of value-of-time when real and hypothetical choices concern time allocation decisions at different occasions. However, we find no evidence of hypothetical bias for a task to be done at once. Moreover, at least in the Chinese sample, we find that ex-post mitigation of negative hypothetical bias by certainty calibration, through re-coding of uncertain "yes" responses into "no", gives rise to another bias, with opposite sign, while calibration by restricting estimations to confident "yes" and "no" responses reduces the bias.

Keywords: Stated choice, Certainty calibration, Preference certainty

JEL codes: C20, C91, D80

* Corresponding author: Lars Hultkrantz, Örebro University, SE-701 82 Örebro, Sweden. +46-90-301416, lars.hultkrantz@oru.se

1. *Introduction*

Hypothetical bias is the main validity concern to the use of survey methods of valuation of non-market price goods. The prevalence of a positive bias that leads to overestimation of the actual (real) value of public and private goods in hypothetical valuation surveys is well documented (Harrison & Ruthström 2008, Murphy et al. 2005), mostly by comparisons of hypothetical choice experiments to economic experiments involving choice that results in a real economic transaction. The degree of the bias varies considerably and seems to depend on context; and cannot be eliminated by a rule-of-thumb approach such as division with some specific number. Resolving this problem is therefore essential in making survey-based methods valid tools for practical use, such as estimation of value parameters in cost-benefit models for evaluation of investments and policies in transport, health, environmental protection, and other sectors. The place to do this is in the economic experiment laboratory; as pointed out by Harrison (2006, p. 127), "although a major thrust of previous experimental work has been directed at undermining the false confidence that hypothetical survey proponents assert in their method, the next stage in this research is likely to emphasise the complementary nature of field and lab valuation exercises".

A special case is valuation of time, for which two studies surprisingly have found evidence of negative hypothetical bias (Brownstone and Small 2006, Isacsson 2008), i.e., the value of time revealed by real choice is higher than the value that is estimated from a hypothetical choice survey. This finding is interesting in itself, as valuation of time (for instance, commuters‘ willingness to pay for a new road that reduces travel time or their willingness to accept to go by a slower travel mode) is an important application of valuation methods. However, it also provides a clue to the explanation of hypothetical bias as such. In a normal choice setting, respondents face a financial budget constraint, while in a time allocation choice they also have to consider a time constraint. Because of a focus effect on the item being valued the financial constraint may be overlooked, which leads to overvaluation; but if the item is a time saving and the time constraint is neglected, under-valuation may result (i.e., respondents do not fully consider alternative, possibly more valuable, ways of using the time, so the opportunity cost of time is underestimated) .

Several remedies to hypothetical bias have been suggested. These include ex ante measures affecting how data is collected and ex post measures affecting the analysis of data. Some ex ante countermeasures, such as making respondents aware of budget constraints, are now standard procedure and others, such as so-called cheap talk scripts, have been tried with mixed results (Cummings and Taylor 1999, Little and Berrens 2004). Recently, Hensher (2009) has suggested that referencing, or pivot design, is another candidate for reduction or even elimination of hypothetical bias. In such design of a value-of-time study, respondents are interviewed about the attributes (such as time and cost) of a reference trip that they already have made (or are undertaking), and are then asked to participate in a stated choice experiment that includes the reference trip as one choice alternative (out of two or three alternatives). In this way, it may be possible to capture context-dependent information about the (dis-)utility of these attributes (and in particular on the actual opportunity costs of time and money), and therefore to get rid of hypothetical bias. Hensher (2009) provides evidence in support of this claim from comparisons of revealed preference and stated preference studies performed in Australia and New Zealand. However, this is not supported by the value-of-time experiments made by Isacsson (2008) where a negative hypothetical bias is held in a setting where respondents were given hypothetical or real choices at the same occasion (but in different rooms).

Ex-post mitigation tries to remove the hypothetical bias by statistical calibration. One such approach is “re-coding”, i.e., change of no-responses to yes for respondents who reveal some degree of less than full confidence to their statement in a follow-up question (Johannesson et al. 1998, 1999, Blumenschein et al. 1998, 2008, Blomquist et al. 2009, Hedemark Lundhede et al. 2009). Another approach is “restricting”, that is, estimation based on the sub-sample of fully confident (or close to fully confident) respondents (Hultkrantz et al. 2006, Svensson 2009 and Sund 2009).

This second alternative has some resemblance with inclusion of a “don’t know” choice option. Both approaches are based on a presumption that responses from individuals who are more certain about their stated responses (intentions) are better predictors of real behaviour, but diverge on how to treat uncertain responses. “Re-coding” assumes that uncertain yes-responses are false (while uncertain no-responses are always true), while the “restricting” method ignores all uncertain (yes and no) responses.[1] While “re-coding” is necessarily conservative in the sense that the share of yes-responses is decreased, “restricting” can go in either way.

A much discussed issue in the literature on ex-post calibration is on where to draw the line between certain and uncertain respondents and how to design the follow-up question (Blomquist et al. 2009). A related but less recognised issue is to what extent answers to such a question really measures preference uncertainty.[2] Possibly, some individuals always are more confident than others, irrespective of whether they made hypothetical or real choices. Furthermore, it could be that respondents report the degree of cognitive effort they used to answer the choice question, instead of the strength of belief that they would actually do as they have stated.

In this paper we use a time-valuation (purchase of time) questionnaire as a test bed for experiments to study the performance of various remedies to hypothetical bias. Our respondents were students at two universities, one in Sweden and one in China. A first motive for our study is that time valuation is one of the most, or perhaps the most, important applications of survey-based valuation, guiding infrastructure planners world-wide. In spite of that, studies of the hypothetical-bias issue in this context are rare. A second motive is that the possibility of a negative hypothetical

[1] Intermediate approaches are possible in which yes and/no responses are weighted differently depending on the degree of self-reported confidence. The Asymmetric Uncertainty Model (ASUM) multiplies the Yes(=1)/No (=0) variable by the certainty score (scaled 0 – 1). The Symmetric Uncertainty Model (SUM) instead adds or subtracts the certainty score to the Yes/No variable, which therefore makes a 0.5 confident Yes equal to a =.5 confident No.

[2] In a review of empirical studies on preference uncertainty in contingent valuation, Akter et al. (2008) reflect in a final remark that “the fundamental issue that needs to be addressed at this point in the development of preference uncertainty research is whether or nor respondent uncertainty can be measured accurately”.

bias provides special challenges to calibration. As far as we know, calibration in such a case has not been studied in any previous work. A third motive comes from the possibility we had to study these issues in two different cultural contexts (Sweden and China) and thus to search for regularities across countries.

Four results from these experiments stand out. First, we find no evidence of hypothetical bias when respondents are randomly given a hypothetical or a real surprise offer for purchase of time at once, hence when the reference situation (that may determine which alternative uses of time that may be considered) is equal to both group of respondents. Second, our results indicate that hypothetical bias in elicitation of value-of-time is negative when the hypothetical choice relates to an action (sacrifice of time) in the future. Third, at least in one of our samples, we find that “re-coding” calibration tends to overshoot (i.e., not just eliminating the negative bias but turning it into a large positive bias), while “restricting” has a mitigating effect. Fourth, we find similar responses to the follow-up question among respondents to both hypothetical and real choice questions.

Next we present the basic theory on how to estimate the value of time, then the experimental design and the results. Finally follows discussion and conclusions.

2. *Value of time estimation*

The value of time is usually defined within the context of household production theory (Becker 1965, DeSerpa 1971). DeSerpa (1971) recognized that time spent in different activities can affect utility in different ways and therefore be associated with different values of time (in Becker´s model there is just one value of time, which is equal to the wage rate). In DeSerpa´s model, an individual maximizes utility given a budget constraint, a time constraint, and a minimum time per (consumption) activity constraint. From the first order conditions, the value of time used for an activity *i,* which we will call the value of time, can be derived as $(\mu - \Psi_i)/\lambda$; where μ is the marginal utility of total time, Ψ_i is the marginal value of the minimum time constraint, and λ is the marginal utility of income. From this we see that the value of saving time in a specific activity may be equal to the overall value, or the opportunity cost, of time μ / λ (which in Becker´s framework is equal to the (after tax) wage per hour), that is when the minimum time constraint is not binding, so Ψ_i is zero. Or it may be less than the overall value, and even negative, when that constraint is binding (in other words, people want to be compensated for spending time on some activities, for instance work, while they may be willing to pay for time on other, more funny, activities).

For a long time, empirical analysis of the value of time was based on the random utility approach developed by McFadden (1974). In this, utility, *U*, is assumed to be linearly dependent on choice attributes, such as cost, C, and time, T. Taking the difference between two alternatives (for instance between two alternative travel modes or routes in an urban road network) we have:

$$dU = \beta_T dT + \beta_C dC$$

With appropriate statistical distribution assumptions, this can be estimated with logit or probit regression, and the value of time can then be calculated from the quota of the regression coefficients β_T / β_C, corresponding to the marginal rate of substitution $(\mu - \Psi_i) / \lambda$ in deSerpa`s model. However, the linear functional form and the statistical distribution assumptions are quite restrictive, so much work has been put into elaboration of the functional form (e.g., Gaudry et al. 1989, Jara-Diaz and Videla 1989, Hensher 1996, Hultkrantz and Mortazavi 2001) and/or distributional assumptions (in recent years, predominantly by development of the mixed-logit model of McFadden and Train 2000).

In spite of such improvement of the random utility model, it remains a problem that the value of time is derived indirectly, from a quota of regression coefficients that is strongly sensitive to model misspecification. During the latest years, value-of-time research (Hultkrantz et al. 1996, Fosgerau 2007, Börjesson 2010) has therefore turned to direct estimation of value of time, which can be done both non-parametrically and parametrically. This approach (originally suggested by Cameron 1986) estimates the willingness to accept (WTA) or willingness to pay (WTP) (or the Hicksian variation) of a time change in "bid space", i.e., from yes or no answers to a price bid (Euro per hour). The simplest parametric specification of such a model, which is used in this study, is a first-order Taylor expansion of (here) the willingness-to-accept function:

$$dU = \alpha_0 + \alpha_1 P$$

where *dU* is the change of indirect utility from accepting the offer and *P* is the price bid. A respondent is assumed to accept the offer when *dU* > 0, hence the ratio $-\frac{\alpha_0}{\alpha_1}$ is the net value of time (the minimum monetary compensation for the sacrifice of an hour).

We further assume that the binary response variable Acceptance (Yes = 1, No = 0) is one when the change of indirect utility plus an i.i.d. error term is positive and zero otherwise. We choose the logistic distribution and estimate this with logit. This is estimated for different subsamples and then the corresponding value of time is calculated for each subsample. Standard errors are calculated with the delta method (first order approximation).

3. Study design

Conjectures

In our study we investigate a number of issues related to hypothetical bias in a value-of-time survey and the performance of various possible remedies to this. First, we want to see whether there is any hypothetical bias when subjects given hypothetical and real WTA choices for time are on equal terms (not separately grouped in different rooms; an equal time-requiring task to be done "here and now"). When all subjects face a similar situation, except for the "genuine" individual variation of the valuation of time (due to individual preferences, scheduling constraints, etc.), we conjecture, in line with the assumption underlying the “referencing” method, that there is no systematic difference in what alternative uses of time and money that subjects consider in hypothetical and real choice. Second, we want to see whether the negative bias result could be replicated. For this, we use a treatment (“Hypothetical (later)”, see below), which was expected to induce hypothetical bias. Third, we wanted to study the performance of the two alternative ex-post calibration methods, i.e., “re-coding” and “restricting” methods (using the numerical scale for self-reported response confidence). Fourth, we wanted to explore differences in responses to this confidence follow-up question in real and hypothetical context.

WTA questionnaire

We asked students in Sweden and China three questions to be answered individually: (1) Would you be willing to perform a quarter-of-an-hour task in exchange for a monetary reward (specified as a certain value, i.e., a bid, that was varied across the subjects; (2) How certain are you about this decision; and (3) Are you male or female. The task was specified as filling in a one of two questionnaires. The given alternative responses to the first question was Yes and No, and to the second question a position on a Likert scale going from 1 (low certainty) to 10 (high certainty).

The first question is a standard discrete choice WTA question. Such questions are regularly asked in value of travel time studies, where for instance a commuter is asked whether she would accept a travel alternative with longer travel time than a reference alternative. Value of travel-time studies are often based on samples of residents or commuters within an age span (for instance residents 18-74 years old) that includes university students, so this sub-population is relevant for a methodological study. However, since our subjects were students at universities where many students live within walking distance from campus, instead of travel we used another time requiring activity that students in both countries are familiar with and were expected to have rather neutral sentiments towards. The design of the study (the first question and the task) resembles the one used by Isacsson (2008); in fact, one of the two questionnaires of our study was borrowed from that study. However, to distract attention from the task as such, we used two different questionnaires for Swedish students (while Isacsson (2008) used only one). Moreover, subjects were given no more information on these questionnaires than that one was about "traffic safety" and the other about "quality of life".

For the Swedish students, we initially used a three-level bid vector at SEK 5, 15, and 30, randomly distributed among the individual students. The mid-level bid was selected to approximate the wage per hour of a simple part-time work. We performed a pilot study in January on a group of teachers students (at the start of their lunch break) using a real setting offering the high price (SEK 30). In the pilot, we noticed that all students accepted this offer and we also found that it was possible for all to fill in one of the questionnaires (both were used) within fifteen minutes. In the final session in November (see below), we skipped the mid-level offer, and thus offered either SEK 5 or SEK 30. For the Chinese students, we used three bid-levels; RMB 1, 2.5, and 5, respectively. Here, the mid-level corresponds to the (hourly) wage rate of an ordinary internship in Shanghai.

The third question on sex was the only question on individual characteristics, because we wanted to keep the survey very short to not miss subjects with a high value of time. Also, respondents were homogenous with respect to age (most were 20-30 years old) and it is difficult to get useful and comparable responses on income from students (some live with their parents, many have seasonal work, etc.), so we did not expect age and income to have much explanatory power as covariates. However, we kept responses from students of different student classes separate so that we in the statistical analysis also could differentiate with respect to class.

Treatments

With this generic value-of-time survey, we investigated three treatments by varying the phrasing of the WTA-question across subjects. First, this question was framed as either a hypothetical choice, with no further consequences, or as a real choice, immediately followed by a fifteen minutes task and monetary reward. In the hypothetical case, subjects were told that they have not been selected for the real task, but asked how they *would have* answered if the question was for real. Second, the hypothetical choice was phrased in two different ways. In one, the task would (hypothetically) be conducted here and now, i.e., as in the real choice situation; and in the other, the task would be done later, at a similar occasion during the semester. Thus, the three treatment are "Real", "Hypothetical (now)", and "Hypothetical (later)".

Procedure

All subjects were recruited without previous notice and in a class-room situation where it was possible to get response from everyone. The teachers had been contacted in advance, but had been asked to not tell the students before or during the lecture. By surprising subjects, we wanted to avoid that results could be affected by scheduling or re-scheduling before or during class.

We came to each class a few minutes before a lecture was over. When the teacher had finished we immediately asked the students, while still seated, to fill in the one-page questionnaire with the three questions. This took at most two minutes. Students that had been given a real offer and had accepted were then asked to stay while other students left the room. Thus, unlike in Isacsson (2008), respondents to both real and hypothetical surveys were in the same room, and therefore similarly exposed to any open or subtle signals from peers and instructors. The remaining participants were then handed one of the two lengthy questionnaires that they had accepted to fill in.[3] Finally, when fifteen minutes had passed, all questionnaires were collected and the students were paid an amount equal to the bid they had been offered individually. In each session, there was a research leader instructing the students. He was assisted by three or four persons so that distribution and collection

[3] In Sweden, students were asked to raise a hand if they did *not* have a drivers license and those that did this were give the Quality of Life questionnaire, while the others got the Traffic Safety questionnaire. In China, where few students have a driver's license, all students were given the former questionnaire.

of questionnaires, and subsequent payments, could be made very quickly, to keep additional time above the stated fifteen minutes at a negligible magnitude.

Subjects and sessions

The WTA survey was conducted in four student classes (263 students) at Örebro University, Sweden, in March 8 and 23, 2009; in three classes(207 students) at Shanghai University of Finance and Economics, China, in May 2009; and finally in another class (98 students) in Örebro in November 2009. All students were in Economics or Business Administration classes. The first Örebro sessions were done in three first-year student classes on March 8 and a third semester class on March 23. On March 8, two classes ended at 3:00 p.m. and the third one at 4:00 p.m. On March 23, the class ended at 3:00 p.m. There were no other scheduled university activities afterwards. The Shanghai s sessions were done in three classes at the beginning of the lunch break, 11:45 a.m. on May 11, 14 and 16. The final Örebro session was done November 9, in a first semester class ending at 3:00 p.m. This session was done after preliminary analysis of the previous results had indicated an outstanding issue (too few observations of fully certain respondents) that possibly could be solved by collection of a larger number of observations.

In the first three classes in Sweden and in China (211 and 207 students, respectively), hypothetical and real choice WTA questionnaires were randomly distributed among the students in the class (107 and 91 students, in Sweden and China respectively in the real choice, and 104 and 116 students, respectively, in the two versions of the hypothetical choice). In the remaining Swedish sessions (52 student March 23, 98 students November 9) only hypothetical choice was surveyed.

4. Results

Estimation method and models

We estimate the logit models on paired “real” and “hypothetical (now)”/ “real “and “hypothetical (later)” sub-samples with a dummy variable indicating observations from each of the hypothetical questionnaires (H=1, otherwise 0) and with the covariates sex and certainty indicator. In addition for the March 8 sessions, we use two dummy variables (Class 2, Class 3) representing the different student classes. Two classes were introduced to the surveys by one of us (the senior author) while the third class (Class 3, ending at the same time as one of the previous) was instructed by a colleague. We use one of the two first classes as reference. For the third wave (China), we use a dummy variable, May 14, with value one for respondents in the sample responding on May 14, zero otherwise. In attempt of eliminating hypothetical bias (α_2 significant) we perform two kinds of certainty calibration: “Re-coding “(changing yes-responses below a specific self-reported confidence level to no-responses) and “restricting “(estimating the model on a restricted set of data, only including Yes and No responses with a confidence level at or above a specific number).

Finally, to explore the responses to the follow-up question we also do regressions (OLS) of the preference certainty variable on various variables.

Descriptive results

In Table 1 we summarize the data, split in six sub-samples (“real” Sweden, “real” China, “hypothetical (now)” Sweden, “hypothetical (later)” – March sessions Sweden, “hypothetical (later)” – November session Sweden, and “hypothetical (later)” China). It can be noted that the distribution of sexes are even in both Sweden and China. The acceptance rate was 0.50 and 0.55 in Sweden and China, respectively, when the offers were for real. This rate is similar (0.51) in the “hypothetical (now)” sub-sample but higher in the “hypothetical (later)” sub-samples (0.71 and 0.65 in Sweden in March and November, respectively, and 0.66 in China). This suggests that we might have a hypothetical bias in the “later” case” but not in the “now” case. Finally, we note that the average certainty does not vary much (from 6.4 to 8.0) between the experiments but is slightly higher in the hypothetical cases than in the real and slightly higher among the Chinese respondents than among the Swedish.

(Insert Table 1 about here)

The value of time

The values of time in these six sub-samples are shown in Table 2. The estimated real case value of time per hour in Sweden is SEK66. This is close to the current after tax minimum wage of work that is common among students as part-time work, for instance work as shop assistant[4], which suggests that the students were regarding the questionnaire responding task as equivalent to such work. However, the corresponding real case value among the Chinese students is RMB 8, i.e., somewhat below the mid-level compensation offer.

[Insert Table 2 about here]

More important to the aim of our study are the differences between these estimated real and hypothetical valuations. For the Swedish data, we see that the sample of respondents that got the "hypothetical (now)" survey on average revealed a value of time of SEK 63 per hour, i.e., just four percent below the value of the real choice survey.

A very different picture is given by the two "hypothetical (later)" samples. The Swedish values in March and November are SEK -412 and -17,respectively, thus considerably lower and indicating that, using de Serpa's framework for interpretation, subjects on average found the joy of performing the task more valuable than the opportunity cost of time. In the Chinese case, the "hypothetical (later)" value of time is a bit more than half of the estimated real value (however, the estimated real value is within the 95 percent confidence interval of this value).

Hypothetical bias

Table 3 shows logit estimates of the random utility model with covariates in three pooled samples: Real and "hypothetical (now)" Sweden; real and "hypothetical (later)" Sweden (all sessions);

[4] The minimum wage per hour is SEK 87.44; after tax (marginal tax rate 26%) SEK 64.71,

and real and “hypothetical “(later)” China. The results are qualitatively similar, with one exception. In all equations, the coefficient of the bid price is positive and significant, while the certainty and sex variables have no significant effect. [5] In the first pooled sample, based on the March 8 sessions, there is a significant effect of Class 3. Since Class 3 was instructed by another researcher than Classes 1 and 2, we interpret this as an instructor effect. In China, the instructor was the same through all experiments, and the dummy variable for one of the classes (May 14) given real choices does not reveal any similar problem. Finally, and most interesting, we find no significant hypothetical bias when using data from the Swedish “hypothetical (now)” sample (see the first column of regression coefficients in Table 3), while we do find a significant bias in the “hypothetical (later)” sample (second column). In the Chinese case, however, the coefficient of this variable is not significant (third column).

[Insert Table 3 about here]

We now turn to ex-post calibration to mitigate hypothetical bias in the “hypothetical (later)” treatment samples in Sweden and China, respectively. In Table 4 we present the estimated value of time the “hypothetical (later)” samples for Sweden and China, respectively, using “re-coding” (i.e., changing uncertain Yes to No) and “restricting” (i.e., not using uncertain Yes and No responses in the estimations) at various cut-off levels. These values are compared to the value of the “real” case (first row).

Table 4 reveals that “re-coding” malfunctions in the (larger) Chinese sample. The fully confident (certainty level = 10) “re-coding” results in a five times higher value of time than the estimated real value, although the initial (full sample) results indicated a negative (although not significant) bias. In the Swedish sample, “re-coding” reduces the magnitude of the negative value but never turns it over to the positive side.

[5] The certainty and sex variables were used as covariates for exploratory reasons only, we had not theoretical expectations for these two variables

“Restricting”, on the other hand, performs pretty good in the Chinese sample. The fully confident (certainty level = 10) “restricting” results in a value of time estimate just 7 percent below the estimated real value. For the Swedish data, however, the outcome of this method is more or less like that of “re-coding”. In the March sessions sub-sample, the magnitude of the negative value is reduced (until the method breaks down at level 9 or 10, probably because of too few remaining observations), but the sign is not changed. In the slightly larger November session sample, “restricting” yields positive value of time estimates from a certainty level cut-off at 8 and higher, but these results are still far below the estimated real value.

Table 4 here.

Explaining preference uncertainty

The results of the follow-up questions under the three Swedish and two Chinese treatments, respectively, are displayed in Figures 1 and 2. As we have already observed, respondents to real choices do also report that they were uncertain to some degree on the choices they made, and in fact they are so more than respondents to the hypothetical choice questions.

Separate OLS regressions[6] of the preference uncertainty variable from the pooled Swedish and Chinese samples, respectively, confirm that the difference in average response between each of the two hypothetical versions of the WTA question and the real case version in Sweden are significant (at the 5 percent level), while the corresponding difference in the Chinese pooled sample is not, when price, sex and acceptance response are used as covariates. None of the covariates are significant. Further exploring this finding by introducing interaction terms, we find one such weakly significant

[6] Stata outprints chan be held from the authors.

(p-value = 0.07) coefficient in a negative coefficient between the price variable and the dummy variable representing the “hypothetical (later)” sample.

5. Discussion and conclusions

In this study, we have followed the Harrison (2006) creed for using lab experiments (with students) to explore ways to mitigate hypothetical bias in value-of-time surveys.

In digestion of the results, we begin to notice that we found no hypothetical bias when respondents to a real and hypothetical value-of-time WTA question were in an equal choice situation because both the real and the hypothetical task was to be conducted at once and at the same location. The sessions were made at the end of the day when most students leave the university. Subjects were not informed in advance and some of them wanted to rise from their seats in the second after the teacher had finished and before we had asked them to fill in the first questionnaire. When they finally left some seemed to be on rush. This indicates that there were individual differences in the value of time. However, when subjects considered their alternative uses of the time that was asked for, by our design there should have been no systematic differences between subjects responding to a real or a hypothetical choice question as all had to consider doing the task immediately and at the same location. Moreover, respondents asked to consider a hypothetical choice to be made at a later occasion answered differently. Our findings therefore confirm that a major challenge in designing a hypothetical choice survey is to ascertain that respondents do not systematically consider a different choice set than that of a real choice. The “referencing” design of choice experiment tries to do this by asking subjects that are in a real choice situation to consider additional hypothetical choice alternatives. According to Hensher (2009) such a design “can deliver the relevant market information as well as attribute variability, while avoiding the problems in identifying meaningful data on non-chosen alternatives”.

However, our results are in contrast to the finding of Isacsson (2008) who in one of his experiments, and the one that resembles ours, did find a negative hypothetical bias. We have no explanation for this difference,[7] but an obvious candidate is that in Isacsson´s experiment, unlike ours, the students were split in separate rooms, one in which all were given a real choice and another where all were given a hypothetical choice. It could be that it was more easy for subjects who were facing a hypothetical choice in our experiment to imagine that it *could had been* a real choice, given that they knew that other subjects (in the same room) were having such a real choice, while subjects given a hypothetical choice in Isacsson´s experiment did not have such a reference. If this is so, this means that ours experiment was a little more close to the "referencing" design (which uses a real trip choice made by the respondent herself as reference).

A second noticeable finding from our experiments is that we, as Brownstone and Small (2005) and Isacsson (2008), do find a *negative* hypothetical bias in stated hypothetical choice related to time. Both these studies indicate an equal relative magnitude of the bias, doubling the VOT, and that is what we get in the Chinese experiment. However, the bias in the Swedish experiments was far off from that magnitude. This should not be of any surprise, as the literature on hypothetical bias in general stresses that the size of the bias is very sensitive to context. Indeed, variability is a main reason for the calibration research, since it is not possible to apply any simple "rule-of-thumb" adjustment method.

Third, our results do not confirm that "re-coding" calibration works. This is a little surprising since several recent studies on the contrary have been very encouraging in support of this approach[8]. However, because of the negative sign of hypothetical bias in our study, we have put "re-coding" into

[7] This objection is not noticed in Hensher (2009).

[8] But several studies have been critical; see especially Akter et al. (2008) and Hedemark Lundhede et al. (2009).

a tougher test than performed by these previous studies. When hypothetical bias is positive, "re-coding" of "yes" responses to a willingness-to-pay question into "no" responses necessarily reduces the bias. But in the context we study, this is not so. On the contrary, on the Chinese data, re-coding invokes a positive hypothetical bias. In fact, overshooting has been observed in previous studies in cases with a positive hypothetical bias (see for instance Johannesson et al. 1998).

Fourth, our application of "restricting" calibration performs better on the Chinese data, but is not convincing on the Swedish data. However, the results from the Chinese samples indicate that more studies using this approach may be worthwhile.

Fifth, and finally, our findings suggest that the preference-uncertainty measure lacks validity. While it is intended to measure the strength of the subject's confidence in her intention to act as stated, we find that subjects given a hypothetical choice scored higher (more confident) than subjects given a real choice on a closely similar follow-up question. Also, frequency distributions were similar. This indicates that answers to this question, which today is commonly used in contingent valuation studies, largely reflects the subject's perceived cognitive effort of making a decision. That effort is likely to be larger in a real choice situation than in a hypothetical. Thus it may be time to take this issue back to the blue-print stage and try to design follow-up questions that can work better as predictors of the step from stated intention to actual behavior!

The main conclusion for value-of-time research from our study is that hypothetical bias should be as much a concern to designers of surveys and users of results from these surveys as it already is in other fields of non-market valuation, in particular environmental economics. Interestingly, however, our results suggests that the "referencing" design that has developed as a best-practice approach in the valuation of time literature, but has not been evaluated previously with economic experiments, could be an effective remedy to hypothetical bias.

References

Akter, S., Bennet J., Akhter S., 2008, Preference uncertainty in contingent valuation. *Ecological Economics* 67:345-351.

Becker, S.G., 1965, A theory of the allocation of time. *Economic Journal* 75, 493-517.

Blomquist G.C., Blumenschein K., Johannesson M., 2009, Eliciting willingness to pay without bias using follow-up certainty statements: comparisons between probably/definitely and a 10-point certainty scale. *Environmental and Resource Economics* 43:473-502.

Blumenschein K., Johannesson M., Blomquist G.C., Liljas B., O'Conor R.M. , 1998, Experimental results on expressed certainty and hypothetical bias in contingent valuation. *Southern Economic Journal* 65:169-177.2.

Blumenschein, K., G. C. Blomquist, Magnus Johannesson, Nancy Horn and Patricia Freeman, 2008), Eliciting willingness to pay without bias: evidence from a field experiment. *The Economic Journal* 118 (January), 114–137.

Brownstone, D. and K. A. Small, 2005, Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part* A 39,279–293.

Börjesson, M. , 2010, Swedish Values of Travel Time and their application in appraisal. Centre of Transport Studies, Royal Institute of Technology, Stockholm, Working Paper.

Cameron, T.A. "A New Paradigm for Valuing Non-Market Goods using Referendum Data: Maximum Likelihood Estimation by Censored Logistic Regression, 1988," *Journal of Environmental Economics and Management*, 15. 355-379.

Champ, P. A. and R. C. Bishop , 2001), Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias. *Environmental and Resource Economics* 19(4), 383-402.

Cummings, R. , Taylor,L.O., 1999, Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method. *American Economic Review*, 89(3), 649-665.

deSerpa, A.C., 1971, A theory of the economics of time. Economic Journal 81, 828-845.

Fosgerau, M., 2007, Using nonparametrics to specify a model to measure the value of travel time. *Transportation Research Part A*, 41, 842-856.

Gaudry, M., S.R. Jara-Diaz, and J. De D. Ortúzar, 1989, Value of Time Sensitivity to Model Specification. *Transportation Research* 23B, 151-58.

Harrison, G.W., 2006, Experimental Evidence on Alternative Environmental Valuation Methods. *Environmental and Resource Economics* 34, 125-62.

Harrison, G.W. and E Ruthström, 2008, Experimental evidence on the existence of hypothetical bias in value elicitation methods. In Plott, C. and V.L. Smith (eds). Handbook of experimental economics results. Elsevier Science, New York.

Hedemark Lundhede, T., Olsen, S.B., Jacobsen, J.B., and Jellesmark thorsen, B., 2009, Handling respondent uncertainty in Choice Experiments: Evaluating recoding approaches against explicit modeling of undertainty. *Journal of Choice Modelling*, 2(2), 118-147.

Hensher, D. A., 2009, Hypothetical bias, choice experiments and willingness to pay. Forthcoming, Special Issue of *Transportation Research B*.

Hultkrantz, L., G.Lindberg, and C. Andersson, 2006), The value of improved road safety. *Journal of Risk and Uncertainty* 32, 151-170.

Hultkrantz, L., C. Li, and G.Lindberg, 1996, 'Some Problems in the Consumer Preference Approach to Multimodal Transport Planning'. Presented at the conference *Transport and Environment*, FEEM, Venice, November 9-10, 1995. CTS Working Paper 1996:5.

Hultkrantz L., and R. Mortazavi, 2001, Anomalies in the Value of Travel-Time Change. *Journal of Transport Economics and Policy* 35, Part 2, 285-300.

Isacsson, G, 2008, The trade off between time and money: Is there a difference between real and hypothetical choices? Swedish National Road and Transport Research Institute
Transport Economics Research Unit (available at the Scandinavian Working Papers in Economics web site).

Jara-Diaz, S.R, and J. Videla, 1989, Detection of Income Effect in Mode Choice. *Transportation Research*23B, 393-400.

Johannesson M, Liljas B, Johansson P-O, 1998, An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. *Applied Economics* 1998;30:643-647.

Johannesson M, Blomquist GC, Blumenschein K, Johansson P-O, Liljas B, O'Conor RM., 1999, Calibrating hypothetical willingness to pay responses. *Journal of Risk and* Uncertainty 18:21-3

Kagel, J. H..Roth, A.E., 1995, *The handbook of experimental economics*. Princeton: Princeton University Press.

List, J. A. and C. A. Gallet, 2001, What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?. *Environmental and Resource Economics* 20(3), 241-254.

List, J. A. and J.F. Shogren ,1998, Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization* 37, 193-205

Louviere, J. J., D. A. Hensher and J. D. Swait, 2002, *Stated Choice Methods: Analysis and Application.* Cambridge: Cambridge University Press.

McFadden, D., 1974, The Measurement of Urban Travel Demand. *Journal of Public Economics* 3, 303-28.

McFadden, D. and K. Train, 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15(5), 447-70.

Murphy, J. J., P. G. Allen, T. H. Stevens and W. Darryl W., 2005, A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics* 30, 313–325.

Swärdh, Jan-Erik ,2009, Hypothetical bias and certainty calibration in a value of time experiment. Örebro University, dissertation.

Svensson, M. , 2009, The Value of a Statistical Life in Sweden Estimates from Two Studies using the "Certainty Approach" Calibration, *Accident Analysis and Prevention*, forthcoming.

Sund, B., 2009, Certainty calibration in contingent valuation - exploring the within-difference between dichotomous choice and open-ended answers as a certainty measure. Örebro University, Dept of Economics Working Paper 2009:1.

Table 1. Descriptive summary of the data (averages with standard errors in parentheses) split in sex sub-samples (“real” Sweden, “hypothetical(now)” Sweden, “hypothetical (later)” Sweden March sessions, “hypothetical (later)” Sweden November session, “real” China”, and “hypothetical (later)” China.

Variable	**Real**	**Hypo**	**Hypo**	**Hypo**	**Real**	**Hypo**	**Min**	**Max**
		Now	**Later**	**Later (Nov)**		**Later**		
	Sw	**Sw**	**Sw**	**Sw**	**Ch**	**Ch**		
Price*	16.58	16.39	17.5	17.5	16.38	16.86	5	30
	(10.38)	(10.25)	(12.62)	(12.56)	(10.26)	(9.90)		
Female	0.40	0.50	0.54	0.55	0.53	0.53	0	1
	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	0	1
Certainty	6.41	7.15	7.29	7.03	7.71	8.03	1	10
	(2.31)	(2.23)	(2.10)	(2.35)	(2.23)	(2.00)		
Accept	0.50	0.51	0.71	0.65	0.55	0.66	0	1
	(0.50)	(0.50)	(0.46)	(0.48)	(0.50)	(0.47)		
Obs	107	104	52	98	116	91		

* Because the price bid vector in Sweden was 6 times the pride bid vector in China, the Chinese price variable has been multiplied by 6 to simplify comparison.

Table 2. Net value of time (SEK/hour and RMB/hour), standard errors in parentheses.

	Real	**"Later"**	**”Later” (Nov.)**	**"Now"**
Sweden	66.13	-412.27	-16.88	63.39
	(16,88)	(1578.43)	(22.32)	(19.16)
China	7.59	4.16		
	(3.73)	(2.58)		

Note: Net VOT is calculated from estimations of logit models (without covariates) on each separate sample. Standard errors are computed by the delta method.

Table 3. Logit model of responses in experiments 1. “Real” and “hypotetical (now)”, Sweden, 2. “real” and “hypothetical (later)” Sweden – all sessions, and 3. “real” and “hypothetical (later)” China. Standard errors in parentheses.

Variable	**Real and “hypoth. (now)”, Sweden**	**Real and “hypoth (later)”, Sweden**	**Real and “hypoth. (later)”, China**
Offer price	0.048***	0.038**	0.344***
	(0.014)	(0.016)	(0.094)
Hypothetical	0.028	0.776**	0.422
	(0.295)	(0.379)	(0.371)
Female	-0.096	0.398	-0.267
	(0.295)	(0.334)	(0.301)
Certainty	-0.402	0.033	-0.051
	(0.065)	(0.075)	(0.072)
Class 2	0.184	-	-
	(0.344)		
Class 3	0.635*	-	-
	(0.379)		
May 14	-	-	-0.171
			(0.441)
Intercept	-0.567	-0.952	-0.098
	(0.570)	(0.627)	(0.701)
No. of observations	204	160	207
Pseudo R2	0.045	0.060	0.067

Note: Significant for $\alpha = 0.01{*}{*}{*}, \alpha = 0.05{*}{*}$.

Table 4. Certainty calibration of the “hypothetical (later)” samples in Sweden (March sessions), Sweden (November session) and China, respectively, using the “re-coding” and “restricting” methods, respectively. First column denotes results based on real sample and hypothetical sample, at different cut-off points of the “certainty” variable, respectively. Standard errors within parentheses (delta method).

“Recoding”

	Sweden (March) VOT (SEK/hour)	No. obs.	Sweden (Nov.) VOT (SEK/hour)	No. obs.	China VOT (RMB/hour)	No. obs
Estimated real	66				7.6	
≥5	-344 (1404)	52	-29.8 (22)	98	4.3 (2.8)	116
≥6	-20.0 (254)	52	-66.6 (15.8)	98	7.3 (2.9)	116
≥7	-20.0 (254)	52	-81.3 (18.2)	98	10.2 (2.6)	116
≥8	-41.9 (139)	52	-159 (80.6)	98	17.1 (5.9)	116
≥9	-76.2 (101)	52	478 (1049)	98	27.3 (9.0)	116
10	-12.0 (111)	52	265 (344)	98	38.8 (17.9)	116

“Restricting”

	Sweden (March) VOT (SEK/hour)	No. obs.	Sweden (Nov.) VOT (SEK/hour)	No. obs.	China VOT (RMB/hour)	No. obs
Estimated real	66				7.6	
≥5	-198 (496)	49	-2.2 (22.7))	82	3.2 (3.1)	198
≥6	-59 (151)	37	-14.5 (16.5)	68	3.4 (3.4)	172
≥7	-59 (151)	37	-13.8 (18.0)	63	3.4 (3.6)	156
≥8	-8 (42)	37	10.9 (38.8)	51	4.1 (3.5)	136
≥9	-24 (129)	16	14.6 (69.7)	27	6.3 (3.8)	97
10	-∞	9	8.6 (140	19	7.1 (5.5)	63

Figure 1 Frequency of responses to the certainty follow-up question, Swedish samples

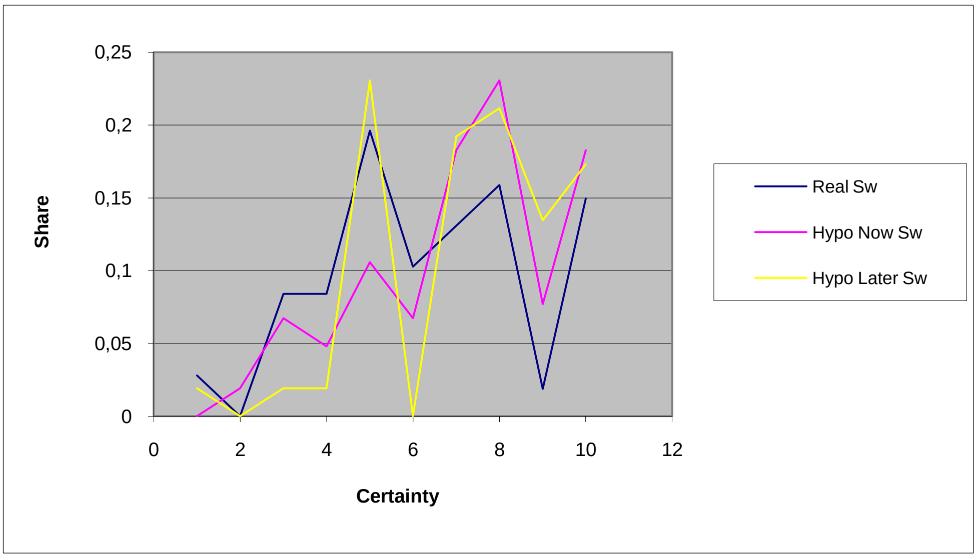
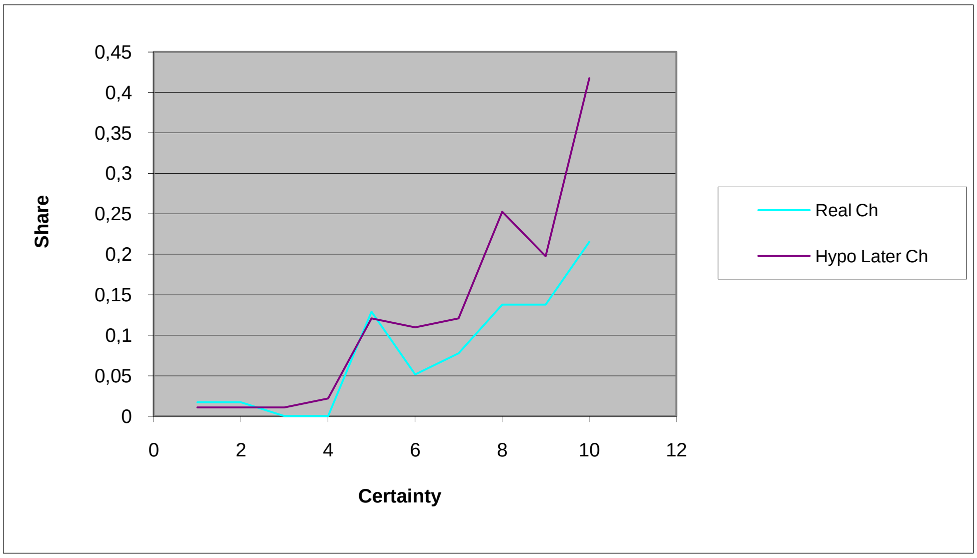


Figure 2 Frequency of responses to the certainty follow-up question, Chinese samples


Share
Certainty
Real Ch
Hypo Later Ch